PURDUE UNIVERSITY

Dealing with Discriminatory Data Mining

Chris Clifton

7 November 2016

With my (former) NSF Hat on:

- Secure and Trustworthy Cyberspace: Enabling US-Brazil Collaboration on Cybersecurity Research
  - jems.sbc.org.br/br_us_cybersec2016
  - www.nsf.gov/pubs/2017/nsf17024/nsf17024.jsp
  - Two-page white papers due 12/16/16
  - See also: www.usbrazilsec.org
- *Nothing else today is from NSF…*

2

What’s all the fuss?

What’s all the fuss?
(Angwin, Larson, Mattu, Kirchner ‘16)

Machine Bias

There’s software used across the country to predict future criminals. And it’s biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

ON A SPRING AFTERNOON IN 2014, Brisha Borden was running late to pick up her god-sister from school when she spotted an unlocked kid’s blue Huffy bicycle and a silver Razor scooter. Borden and a friend grabbed the bike and scooter and tried to ride them down the street in the Fort Lauderdale suburb of Coral Springs.

- Similar cases lead to different outcomes
  - Minor theft (shoplifting, stealing a bike)
  - Black offender predicted as more likely to commit future crime than white
  - *Despite white offender having criminal record!*
- Statistical analysis suggests this is common

What's all the fuss? *(Sanburn '15)*

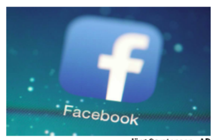
Facebook Thinks Some Native American Names Are Inauthentic

Josh Sanburn @joshsanburn | Feb. 14, 2015

The social network is barring some Native Americans from logging in

If you're Native American, Facebook might think your name is fake.

The social network has a history of telling its users that the names they're attempting to use aren't real. Drag queens and overseas human rights activists, for example, have experienced error messages and problems logging in in the past.



Jörg Carstensen—AP

Some Native Americans say Facebook won't allow them to log in because their names are "inauthentic."

- Ms. Lone Elk (and others) required to provide identification to use Facebook
  - Viewed as potential violation of "real name" policy
- No such barriers for "dominant majority"

What's all the fuss? *(Sweeney '13)*

Discrimination in Online Ad Delivery

Latanya Sweeney
Harvard University
latanya@fas.harvard.edu

January 28, 2013[1]

Abstract

A Google search for a person's name, such as *"Trevon Jones"*, may yield a personalized ad for public records about Trevon that may be neutral, such as *"Looking for Trevon Jones? …"*, or may be suggestive of an arrest record, such as *"Trevon Jones, Arrested?..."*. This writing investigates the delivery of these kinds of ads by Google AdSense using a sample of racially associated names and finds statistically significant discrimination in ad delivery based on searches of 2184

- Blacks and whites see different ads on the internet
  - *Even if race not part of the profile*
- Sweeney found that first names typically associated with blacks and whites lead to different ads
  - Otherwise identical profiles and histories

What's all the fuss? *(Datta,Tschantz, and Datta '15)*

DE GRUYTER OPEN Proceedings on Privacy Enhancing Technologies 2015; 2015 (1):92–112

Amit Datta*, Michael Carl Tschantz, and Anupam Datta

Automated Experiments on Ad Privacy Settings

A Tale of Opacity, Choice, and Discrimination

Abstract: To partly address people's concerns over web tracking, Google has created the Ad Settings webpage to provide information about and some choice over the profiles Google creates on users. We present AdFisher, an automated tool that explores how user behaviors, Google's ads, and Ad Settings interact. AdFisher can run browser-based experiments and analyze data using machine learning and significance tests. Our tool uses a rigorous experimental design and statistical analysis to ensure the statistical soundness of our results. We use AdFisher to find that the Ad Settings was opaque about some features of a user's profile, that it does provide some choice on ads, and that these choices can lead to seemingly discriminatory ads. In particular, we found

serious privacy concern. Colossal amounts of collected data are used, sold, and resold for serving targeted content, notably advertisements, on websites (e.g., [1]). Many websites providing content, such as news, outsource their advertising operations to large third-party ad networks, such as Google's DoubleClick. These networks embed tracking code into webpages across many sites providing the network with a more global view of each user's behaviors.

People are concerned about behavioral marketing on the web (e.g., [2]). To increase transparency and control, Google provides Ad Settings, which is "a Google tool that helps you control the ads you see on Google services and on websites that partner with Google" [3].

- Study of impact of different ad privacy settings
- Disclosing Gender resulted in fewer ads for high-paying jobs

What are the reasons?

- Discrimination programmed into the system?
  - Let's hope not
- Historical bias in the training data?
  - May explain some, but not all
- Insensitivity on the part of developers?
  - Maybe
- Or perhaps we don't know (yet)?

Potential sources

- Historical bias in training data
  - Can we detect this?
- Feedback bias
  - Complexo da Maré has high crime
    - Increase police presence
  - Even more crime discovered in Complexo da Maré
    - Has even worse crime statistics!
- “Tyranny of the majority”
  - Small populations deemed outliers
  - Algorithms effective “on average”, but ignore rare cases
- Wrong objective function
  - Is accuracy the right measure?

What can we do?

- Detect discriminatory outcomes from machine learning
  - [Pedreschi08, Pedreschi09, Luong11, Ruggieri11]
- Relabel training samples
  - [Kamiran09, Zliobaite11, Kamiran11]
- Adjust scoring functions
  - [Calders10, Kamiran10]
- statistical parity
  - [Dwork12, Zemel13]

Disparate Treatment vs. Disparate Impact

- Disparate treatment: Individuals from different groups treated differently
  - Otherwise identical individuals have different outcome based only on group membership
- Disparate impact: Outcomes different between different groups
  - No individuals are "the same"
  - Different outcomes for different groups, even if some other explanation
- Methods on previous slide address disparate treatment
  - But discrimination shows up even when the groups aren't part of the input!

Why Disparate Impact?

- Mortgage Redlining
  - Racial discrimination in home loans prohibited in US
  - Banks drew lines around high risk neighborhoods!!!
  - These were often minority neighborhoods
  - Result: Discrimination (redlining outlawed)

  What about data mining that "singles out" minorities?

Dealing with Disparate Impact
(Mancuhan and Clifton, AI&Law'14)

- Goal: Bayesian classifier that reduces disparate impact on protected group
  - *Group not known when classifying a new instance*
- Idea: Adjust "discriminatory" network
  1. Learn network with protected group known
  2. Identify and relabel victims of disparate treatment
  3. Remove protected group from network
  4. Adjust weights to work with relabeled data

Identifying Discrimination "Victims"

- Assume sets of attributes
  - *p* (protected group membership)
  - *r* (high correlation with protected)
  - *b* (okay to use)
- $belift = \frac{P(C|p_1, p_2, \ldots, p_l, b_1, b_2, \ldots, b_n, r_1, r_2, \ldots, r_n)}{P(C|b_1, b_2, \ldots, b_m)}$

(this is a probabilistic interpretation of the elift definition of Pedreschi et al.)

- *belift* = 1 → no discrimination

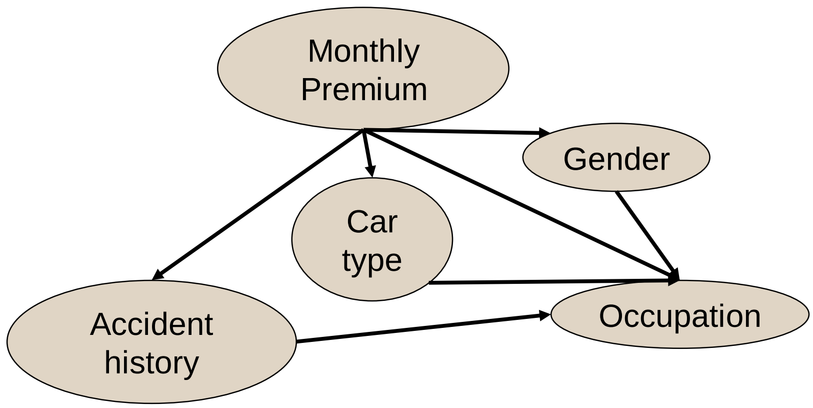
Challenge: Which are the "redlining" attributes

- Distinguishing between *b* and *r*
  - Assume *p* is given
- Build Bayesian network using all attributes
  - Parents and children of protected attributes are presumed to be in *r*
- Remove *p* and *r* nodes to get network to calculate denominator

Build "safe" network

- Identify instances with high *belift*
  - "Flip" class with lowest *belift* to balance distribution in protected groups with overall distribution
- Remove protected attributes from network
  - But keep redlining attributes
- Reweight by training with modified training data
  - Adjusts weight of redlining attributes to avoid use as surrogate for protected attribute

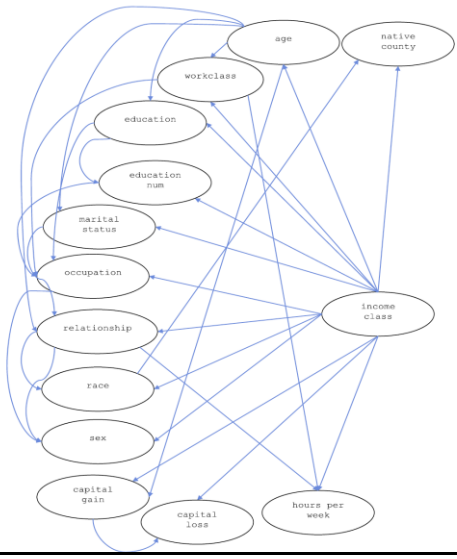
Adjusting the Network



German Credit: Network

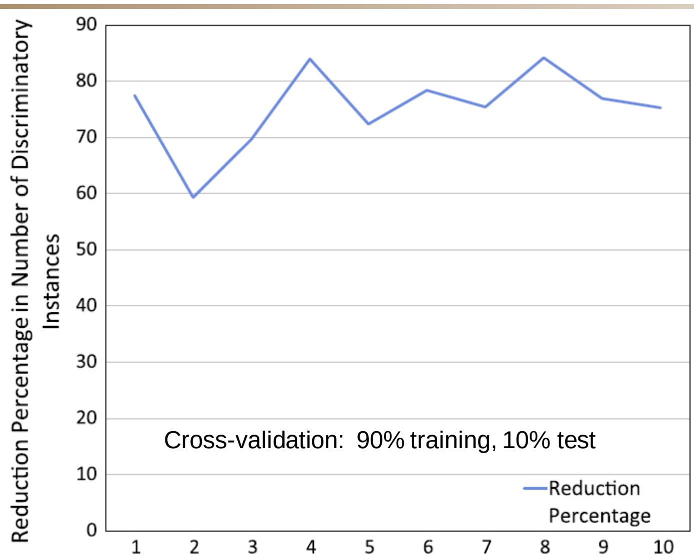
- Little evidence of discrimination
  - Personal_status showed some
  - Correlated with number of dependents
- 10 instances
  - belift values from 1.02-2.75

Census Income: Network

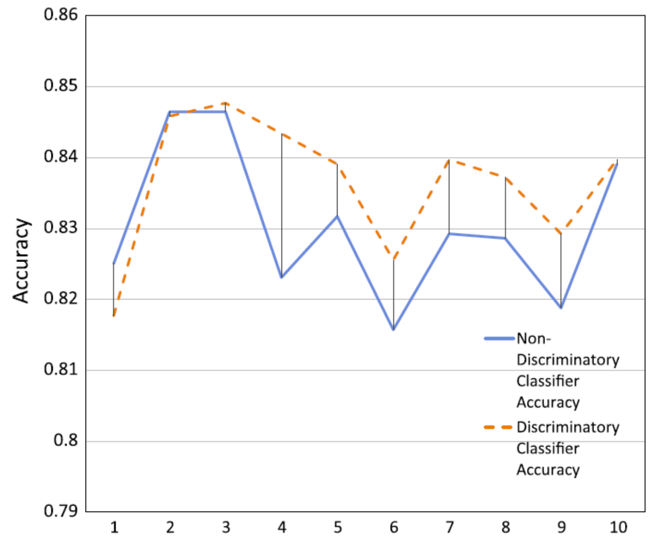


- Some evidence of gender discrimination
  - 688 instances
  - belift median 2, max 20
- Related to
  - Relationship
  - Occupation
  - Marital status

Reduction in Discrimination: Census



Accuracy: Census



Ideas for the Future

- Tests for Bias?
  - Or perhaps just *potential* bias?
- Fundamental changes in machine learning?
  - Objective functions other than accuracy

Ideas? Let's talk!

- Secure and Trustworthy Cyberspace:
  Enabling US-Brazil Collaboration on Cybersecurity Research
  - jems.sbc.org.br/br_us_cybersec2016
  - www.nsf.gov/pubs/2017/nsf17024/nsf17024.jsp
  - Two-page white papers due 12/16/16
  - See also: www.usbrazilsec.org
- Topics:
  1. Security and privacy in networks
  2. The Internet of Things and cyber-physical human systems
  3. Malware detection

23